JOURNAL OF CHEMICAL INFORMATION AND MODELING

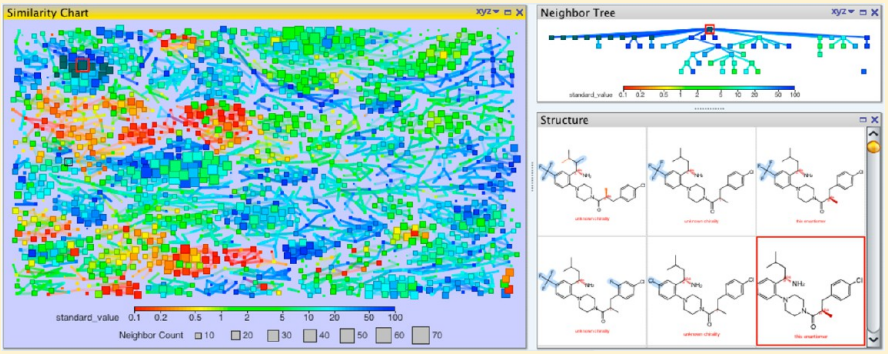
Article

pubs.acs.org/jcim

DataWarrior: An Open-Source Program For Chemistry Aware Data Visualization And Analysis

Thomas Sander,* Joel Freyss, Modest von Korff, and Christian Rufener

Department of Information Management Drug Discovery, Actelion Ltd., Gewerbestrasse 16, CH-4123 Allschwil, Switzerland



ABSTRACT: Drug discovery projects in the pharmaceutical industry accumulate thousands of chemical structures and ten-thousands of data points from a dozen or more biological and pharmacological assays. A sufficient interpretation of the data requires understanding, which molecular families are present, which structural motifs correlate with measured properties, and which tiny structural changes cause large property changes. Data visualization and analysis software with sufficient chemical intelligence to support chemists in this task is rare. In an attempt to contribute to filling the gap, we released our in-house developed chemistry aware data analysis program DataWarrior for free public use. This paper gives an overview of DataWarrior's functionality and architecture. Exemplarily, a new unsupervised, 2-dimensional scaling algorithm is presented, which employs vector-based or nonvector-based descriptors to visualize the chemical or pharmacophore space of even large data sets. DataWarrior uses this method to interactively explore chemical space, activity landscapes, and activity cliffs.

INTRODUCTION

In the pharmaceutical industry as well as in science in general we notice a paradigm shift[1] from hypothesis driven reasoning to data driven approaches. The buzz phrase "Big Data" is touted everywhere and data mining, machine learning,[2] and data visualization[3] are increasingly used to uncover the knowledge hidden in ever growing data sets. An intuitive example that dynamically visualizes the world's most important trends is Hans Rosling's Gapminder.[4] The availability of a multitude of data visualization and analysis software tools is both evidence and driver of the evolution toward data driven science. And yet, while existing software handles multidimensional, alphanumerical data rather well, it is not yet an easy task using software to analyze or visualize the correlation between structural motifs of molecules and alphanumerical data. Nevertheless, proprietary compound and activity databases have grown to an extent that simple search strategies and result table browsing are not anymore adequate methods to extract the relevant information. And with the advent of large, publicly available data sources like PubChem and ChEMBL there is a substantial demand for software tools that seamlessly combine cheminformatics algorithms, physicochemical property prediction, multivariate data analysis, and interactive visualization.

For more than three decades traditional chemistry software companies focused on providing chemical database systems for molecule and reaction storage, search, and retrieval.[5] On the client side they offered chemical editors, form based views, and eventually chemical spreadsheets. Also for 30 years, academic groups and commercial organizations (Tripos, MSI, Schroedinger, CCG) have specialized in modeling software focusing on force field based, semiempirical, and quantum chemical methods for exploring molecule conformations and pharmacophore patterns, predicting compound−protein interactions and potential biological activities. Strong cheminformatics tools have been developed for other specialized purposes, such as combinatorial library design or even synthesis planning.[6,7] QSAR-Software was invented that derives vectorized numerical descriptors from molecular structures and tries to correlate these vectors with molecular properties and biological activities in order to build predictive models. In 1996 the business intelligence tool Spotfire[8] was released and soon used in the pharmaceutical industry for visually exploring molecule related alphanumerical data. In 2007 the Spotfire company was acquired by TIBCO, which 5 years later announced a strategic alliance with PerkinElmer to exclusively sell the Spotfire software in healthcare and other scientific markets. In 2011 PerkinElmer had acquired CambridgeSoft[9] and their widely used chemistry

Received: September 29, 2014

DOI: 10.1021/ci500588j
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

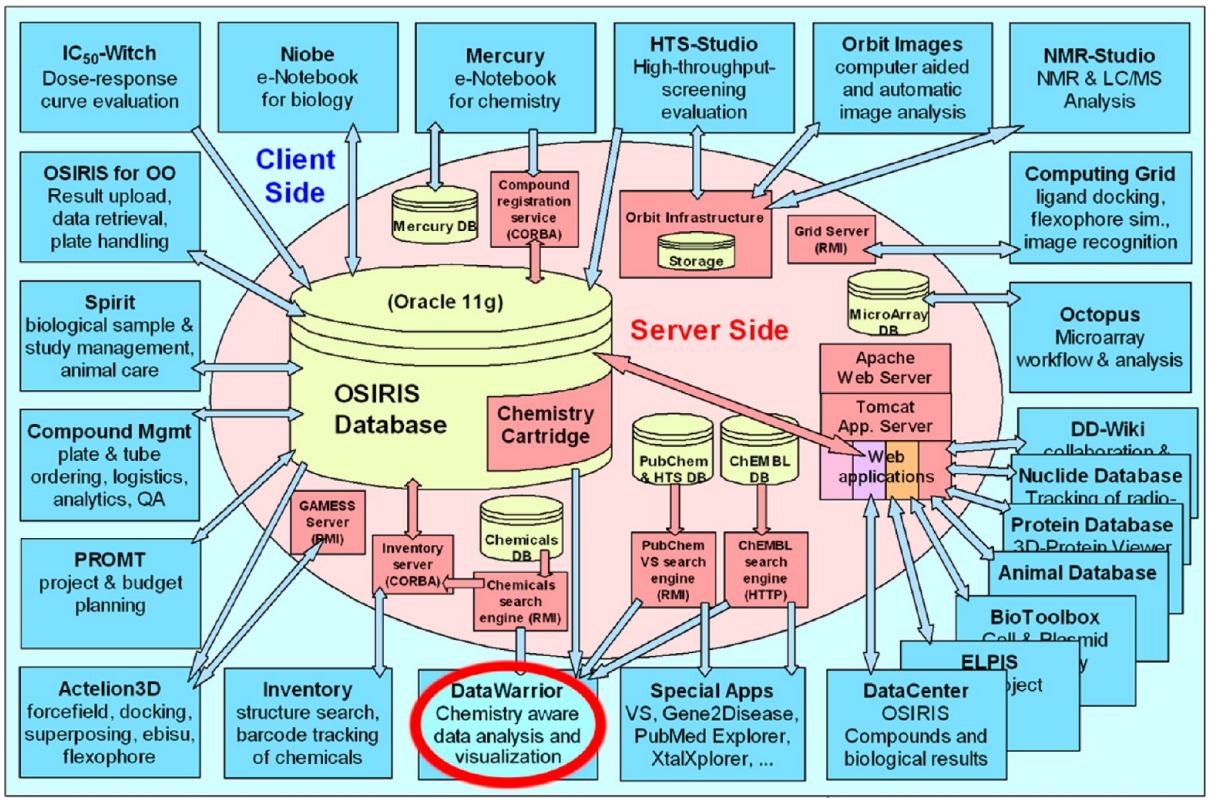


Figure 1. Major components of Actelion's drug discovery application and database landscape OSIRIS.

software products ChemDraw, ChemOffice, and E-Notebook. One goal of this alliance is to more closely connect the former CambridgeSoft software with Spotfire and, thus, to add chemical intelligence to Spotfire. About 10 years after the launch of Spotfire, another company, Dotmatics, launched Vortex,[10] a Java based data visualization software being capable of depicting chemical structures, calculating molecule properties, and clustering molecules from perceived structure similarities. While Spotfire is limited to Microsoft Windows only, Vortex is available for Linux, Macintosh, and Windows.

In 1998, after Actelion Pharmaceuticals Ltd. had been founded, its young drug discovery department faced the challenge to establish from scratch within a few months an information management platform that would support the basic drug discovery workflow including a database for chemical and biological information, a compound registration system to uniquely identify purchased or synthesized molecules, a component to upload biological data from spreadsheets, and a chemistry aware browser for project related data. Furthermore, the chosen architecture should provide a robust and scalable foundation for further growth concerning data and functionality. Components should be chosen wisely to avoid unnecessary dependencies on operating systems or proprietary technologies. It was decided to build the database schema in Oracle, to custom-develop a substance registration system with structure canonicalization in Java, and to develop data retrieval and upload functionality as a Microsoft Excel Add-In. A nightly running process was planned to compile chemical structures and experimental results into project specific ChemFinder[11] files to provide project data browsing with structure search capability. Within a few months Actelion's drug discovery software platform named OSIRIS[12] was operational. In the upcoming years the database structure was extended and new components were developed to embrace most aspects of the drug discovery process. The MS-Excel Add-In was discontinued and replaced by an OpenOffice Plug-In, because OpenOffice can be customized in Java. The ChemFinder file based approach was replaced by a Java application that could directly connect to the database. In parallel an application called DataWarrior was developed to interactively visualize and analyze project data with full support for chemical structures. Fifteen years after its inauguration OSIRIS has grown to a mature, full blown drug discovery platform (Figure 1).

Today DataWarrior combines dynamic graphical views and interactive row filtering with chemical intelligence. Scatter plots, box plots, and bar or pie charts visualize numerical or category data along with chemical information as shared scaffolds and compound substitution patterns. Chemical descriptors independently encode various aspects of chemical structures, e.g. the chemical graph, chemical functionality from a synthetic chemist's point of view or 3-dimensional pharmacophore features.[13] These allow fundamentally different kinds of molecular similarities being calculated on the fly to be used in graphical views or for row filtering and other purposes. Special features support different stages of drug discovery from the selection of lead-like screening compounds through structure activity analysis to the statistical interpretation of animal experiments. DataWarrior supports the enumeration of combinatorial libraries and the generation of evolutionary libraries. Various methods can be applied to visualize chemical space using any of the similarity measures. Physicochemical properties can be calculated and together with other data applied to a multivariate data analysis. Structure activity relationship tables can be created and activity cliffs can be visualized.

DOI: 10.1021/ci500588j
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

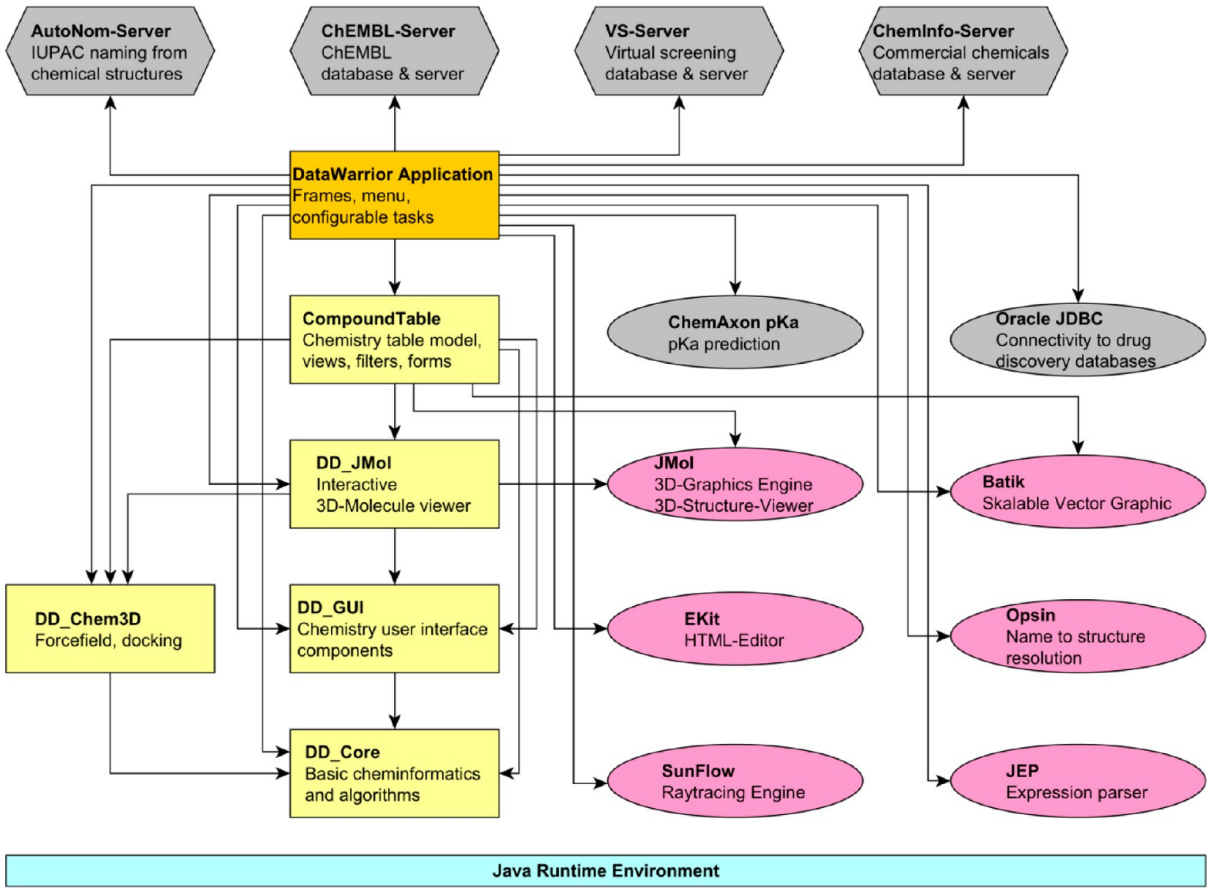


Figure 2. DataWarrior modules and their dependencies. Components that are exclusively used at Actelion and are not part of the public version are drawn in gray.

In the following some of DataWarrior's visualization and analysis features will be demonstrated using chemical and biological sample data.

■ ARCHITECTURE

We had chosen Java as the programming language and framework for the OSIRIS platform, because Java applications run on all major operating systems, the Java runtime environment already includes a rich functionality framework, an active community and commercial vendors provide a multitude of reusable code, and major software components including Oracle databases and OpenOffice can be directly customized and enhanced with Java code.

Thus, when the development of DataWarrior started, we already had built other OSIRIS components with embedded cheminformatics functionality. Therefore, we could already build on a foundation of Java based cheminformatics classes like substructure and similarity searches, molecule depiction, and a molecule editor. Today DataWarrior consists of six in-house developed modules and six open-source libraries. The integrated, nonpublic DataWarrior version used at Actelion also makes use of a commercial module by ChemAxon for pK_a prediction[14] and has functionality to access various in-house databases via HTML interfaces or via a JDBC library provided by Oracle.

All DataWarrior modules of the public version (Figure 2) are described in more detail in the following:

DD_Core. Cheminformatics and other algorithms reused in many OSIRIS projects. Among others these include molecule and reaction classes, chemical file parsers and writers, ring, aromaticity and enhanced stereo perception, structure canonicalization, substructure search, chemical descriptors to calculate multiple aspects of molecular similarity, clustering, substituent and fragment handling, molecule rendering, physicochemical property and toxicity prediction, conformer generation, regression, principal components analysis, and self-organizing maps.

DD_Chem3D. This module contains a force field for molecule energy minimization, a molecule class optimized for proteins and force field operations, the opaque 3D-descriptor Flexophore[13] published earlier, protein and ligand super positioning, structure surface generators, a 3D-structure editor, general functions to handle molecules in 3D-space, and parsers and writers for various file formats of 3D-structures.

DD_GUI. User interface components shared among multiple projects, e.g., a chemical editor, molecule and reaction panels, an editable molecule list, clipboard and drag and drop support, molecule and reaction table cell renderers, a dockable view framework, Windows metafile support, and custom user interface components.

DD_JMol. This module is a wrapper for the JMol library. It contains a 3D-structure editor built on top of the JMol out-of-the-box functionality.

CompoundTable. This module contains a chemistry enhanced table model, where columns may contains numerical, categorical, or text data, chemical structures or reactions, chemical descriptors, or 2D- or 3D-atom coordinates. Columns may be invisible and/or reference another parent column. The table model allows rows to be added, changed, or deleted. It

DOI: 10.1021/ci500588j
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

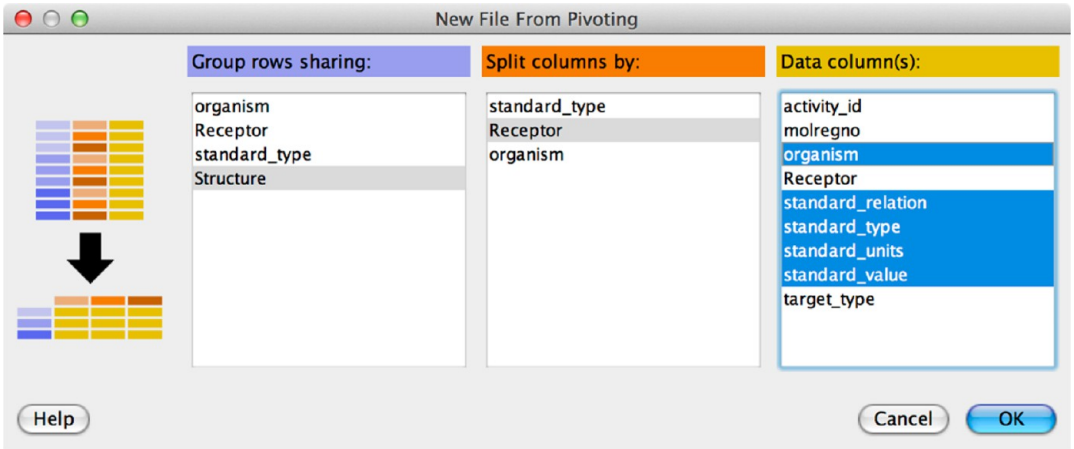


Figure 3. Pivoting Dialog, which is shown when "New From Pivoting..." is selected from the File menu. The dialogue shows three column name lists, everyone populated with those column names, which qualify for a specific role of the pivoting process. After one or more columns are selected in every list, DataWarrior generates a new document by pivoting data from the current window. This process merges data (yellow columns) from many rows into multiple columns, if the data refers to the same object (blue column) but belongs into different categories (orange columns).

handles row selection and sorting, and it allows the user to attach large text or image objects to individual cells. This module contains interactive column type specific filters to modulate row visibility. It also contains 2- and 3-dimensional graphical views, a form based view, a table view, and a chemical structure view. In addition it contains the classes needed to write and read a table model to or from the file system.

DataWarrior. This module constructs all application windows from other module's components. It creates the menu and has implementations for all actions that can be launched from the main menu or the popup menus of the DataWarrior user interface components. It provides dialogues for more complex actions, which need to be configured before being started and it contains some of the algorithms needed to execute these actions. It also contains the functionality to record, modify, execute, and save action sequences as editable and reusable macros.

Batik.[15] Open-source library of the Apache Software Foundation, which is used by DataWarrior for displaying and the creation of Scalable Vector Graphics.

JEP. This is the open-source version 2.4.1 of the Java Math Expression Parser.[16] It is used by DataWarrior to calculate a new column from a definable mathematical expression.

JMol.[17] An open-source library providing a generic 3D-graphics engine with Z-buffer and primitive rendering. It also provides a scriptable interface for 3-dimensional rendering of protein structures, small molecules, and surfaces.

Opsin.[18] An Open Parser for Systematic IUPAC Nomenclature developed by Daniel Lowe at the University of Cambridge as open-source project.

EKit.[19] Open source HTML editor used by DataWarrior to edit file comments.

SunFlow.[20] Raytracing engine written by Christopher Kulla and used by DataWarrior to render photorealistic 3-dimensional molecule images.

FUNCTIONALITY

The DataWarrior application was developed and its functionality evolved over more than 10 years. The purpose of this paper is to give an overview of the functionality available to the public and to discuss a few representative features in more detail. Giving a complete and detailed description of all functionality is beyond the scope of this paper. The interested reader may refer to the DataWarrior online manual.[21]

Data Sources. Internally, the core structure of a DataWarrior document is a data table that contains objects, e.g. molecules, in rows and their properties in columns. A new DataWarrior document can be created by opening a file, reading data from databases or pasting tabular data, e.g. from a spreadsheet application, directly into DataWarrior. In addition to DataWarrior's native file format, DataWarrior reads and writes tab-delimited or comma-separated text files as well as MDL SD-files (Version 2 or 3). If DataWarrior opens a text file with a column title being "Smiles", then it automatically generates chemical structures with proper atom coordinates from the column entries provided that these are valid SMILES[22] codes. New DataWarrior files can also be created by extracting, merging, pivoting, or reverse-pivoting data from existing files. Exemplarily, Figure 3 shows the dialogue for defining pivoting options. Finally, sometimes new DataWarrior files are created as a result of some processing, e.g. when comparing two files for similar compounds or when creating a new combinatorial library.

Views And Filters. When DataWarrior creates a new document from any source, which is not a native DataWarrior file, then it also creates default views and default filters. A mandatory table view shows the data in a natural way. Other default views are a 2- and a 3-dimensional graphical view and, if the data contains chemical structures, a structure view. Views can be added or removed any time. Existing views can be resized or rearranged to coexist side-by-side or stacked on top of each other. The axes of graphical views are initially assigned to those data columns, which correlate best, but can be reassigned to any column containing numerical data or categories. Range sliders allow zooming in on individual axes in order to focus on a data subset. Popup menus allow to customize any view in many aspects. Graphical views may show multiple data dimensions at once with some data columns being assigned to axes, and others being represented by marker sizes, shapes and colors. Labels and chemical structures can be shown on markers or on axis scales. 2D-views can be configured to show a scatter plot, bar chart, pie chart, box plot, or a whisker plot. Moreover, 2D-views can be split into multiple smaller views with each one showing a single data category. Lines can be added to connect markers belonging to the same category or to visualize graph relationships.

DOI: 10.1021/ci500588j
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

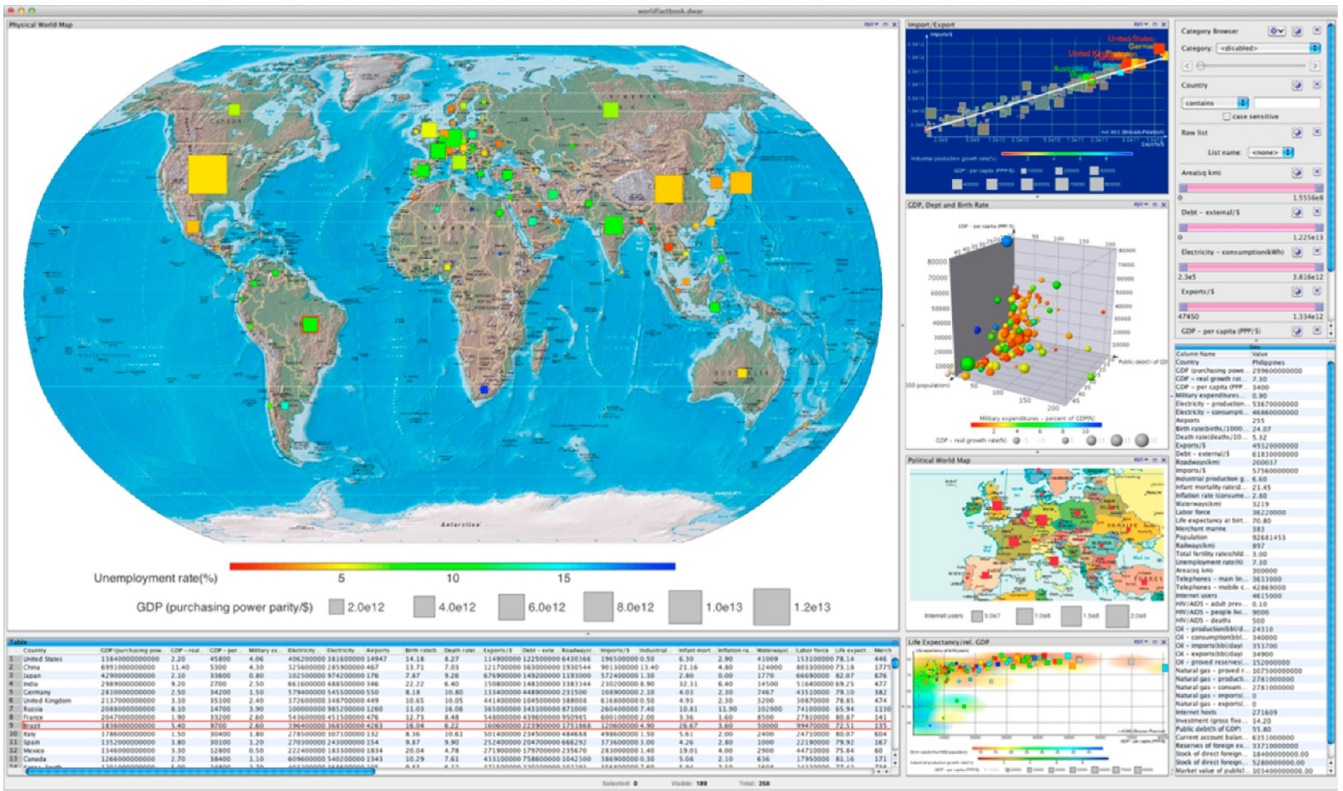


Figure 4. Typical DataWarrior window. Various views show different aspects of the same underlying data table. Data filters (top right) allow the user to explore data subsets. A detail area (bottom right) shows all data of the row in focus.

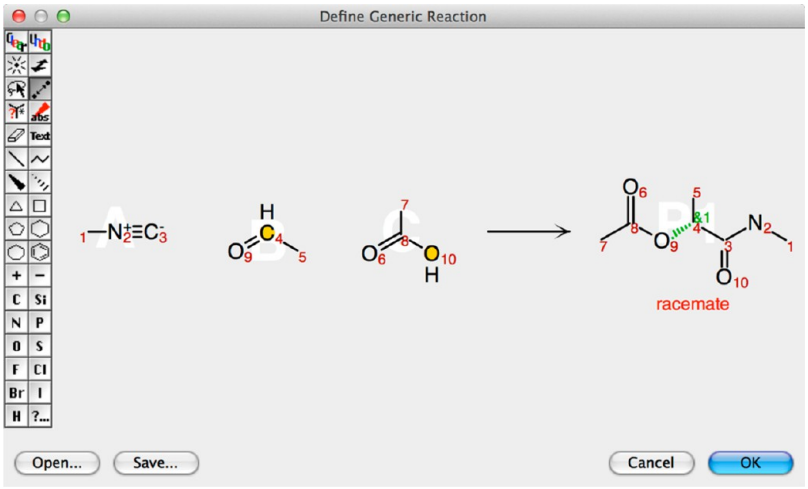


Figure 5. Reaction editor. A Passerini reaction has been drawn, and the reactant atoms have been mapped to product atoms. Yellow markers indicate atom query features. In this example the aldehyde carbon and the single bonded oxygen of the carboxylic acid are defined to fill free valences with hydrogen atoms.

An optional form based view can be added to show all or some data of one data row. Navigation buttons allow stepping back and forth through a data set. Forms are composed of resizable fields showing alphanumerical column data, 2- or 3-dimensional chemical structures, or even attached images or HTML content. A form designer allows customizing the form layout and a form editor allows editing the data shown within the form.

Figure 4 shows a typical DataWarrior window with a table view and five graphical views all showing different aspects of the same underlying data from The World Factbook 2007, released by the CIA.[23]

Data filters, which are always located in the top right corner of a DataWarrior window, are used to interactively hide all data rows from the view that do not match the filter criteria. Filters are usually connected to a data column and the filter kind reflects the column's data type. Range sliders are used to filter numerical data and category filters allow focusing on certain categories. Text filters can be applied to any kind of columns, while structure filters hide rows if a chemical substructure is not present or the similarity to a given structure falls below a definable limit. If rows are assigned to lists, then a list filter may be used to exclusively show list members. Filters can be inverted to hide those rows that

DOI: 10.1021/ci500588j
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

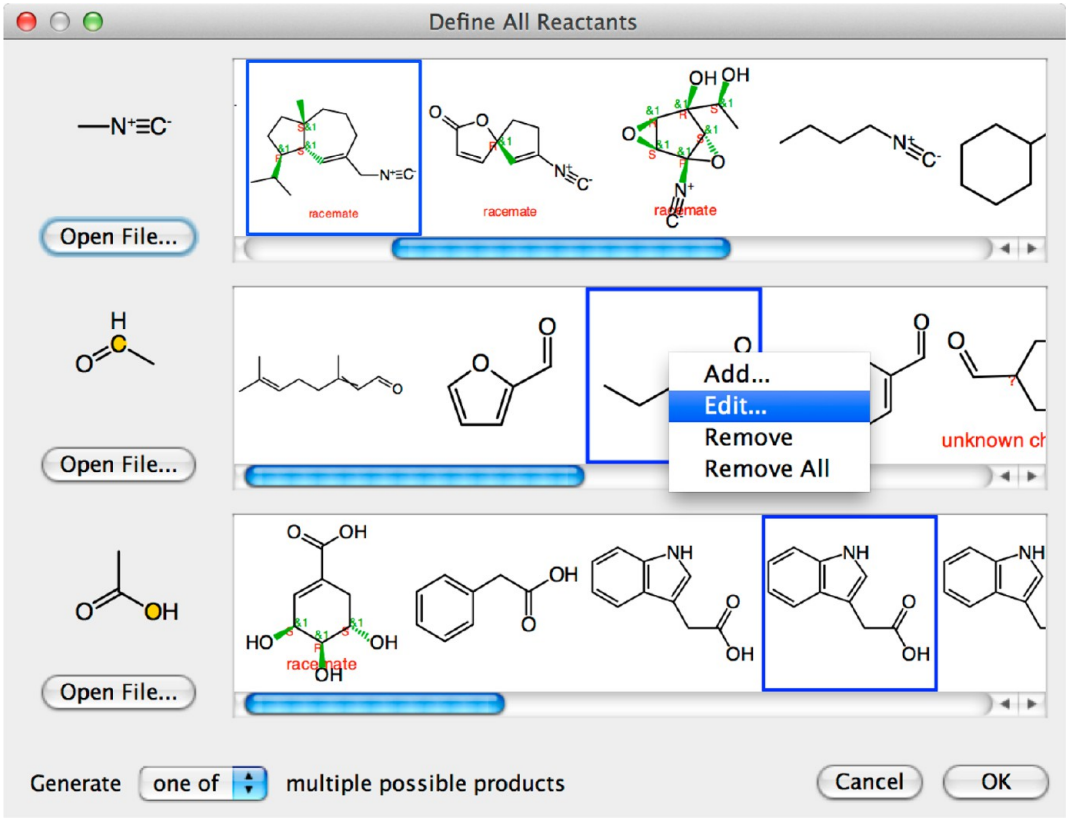


Figure 6. Reactant dialog. Ten compound structures have been defined for each generic reactant. These reactant structures together with the generic Passerini reaction comprise the complete definition of a combinatorial virtual library. A popup menu allows the user add or remove structures from the list or to modify individual structures.

match the filter criteria rather than those that do not match. If multiple filters are used simultaneously, then only those rows remain visible, which are matching all filter's conditions.

Chemical Data. For DataWarrior the chemical structure is a native data type. It can be displayed and used for many purposes including row filtering, customization of graphical views, principal component analysis, or the calculation of self-organizing maps.

When DataWarrior opens a file containing chemical structures such as SD-files, it automatically creates three columns, a parent one for the canonical structure, and two invisible daughter columns for atom coordinates and for a default fingerprint descriptor *FragFp* that is used for fast structure filtering and similarity calculations. Other molecular descriptors can be calculated to support multi purpose similarity measures.

DataWarrior predicts physicochemical and other properties directly from chemical structures. Toxicity risks are predicted from precompiled fragment lists using a previously published algorithm.[24] Dedicated built-in cheminformatics functionality supports various stages of the drug discovery workflow, e.g., an evolutionary algorithm for creating novel structures, virtual screening, compound clustering, structure activity analysis, activity cliff analysis, and many more.

Combinatorial Library Enumeration. In drug discovery combinatorial or parallel chemistry account for a decent percentage of the molecules synthesized. A challenge is to identify and synthesize a diverse and most promising subset of the virtual compound space given by all possible reactant permutations of a multicomponent reaction or multistep reaction sequence. DataWarrior allows to enumerate all possible product structures, predict their properties and to identify diverse subsets of the best candidates. In order to enumerate the product structures of a combinatorial library, the user needs to draw a properly mapped generic reaction as shown in Figure 5. The reaction editor perceives reactants and products from their positions. If disconnected fragments are drawn closely to each other, these are perceived as the same reactant or product, which permits defining a salt as a single compound. Reactants and products are automatically indicated with large white letters as A, B, C and P1, P2, respectively.

After clicking "OK" DataWarrior perceives the number of reactants and opens a dialogue that asks to provide real world compounds for each of the generic reactants (Figure 6). These reactant structures can be imported from a DataWarrior- or SD-file, pasted in from the clipboard, added via drag and drop, or drawn from scratch in the structure editor.

Once the real reactant structures are defined, DataWarrior creates all product structures, which takes about 2 s per 1000 products on a modern desktop computer. Figure 7 shows a DataWarrior window after enumerating 1000 product structures from the Passerini reaction that was used as example in Figures 5 and 6.

Activity Cliff Analysis. One of the biggest challenges of any drug discovery project is understanding the relationship between molecular structures and their biological activities. Of particular interest are the so-called activity cliffs, where small differences in structure cause large changes in biological activity. Exploring these frontier areas of the molecule-activity space provides crucial hints to guide structural modifications toward a compound with an optimized property portfolio.

DOI: 10.1021/ci500588j
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

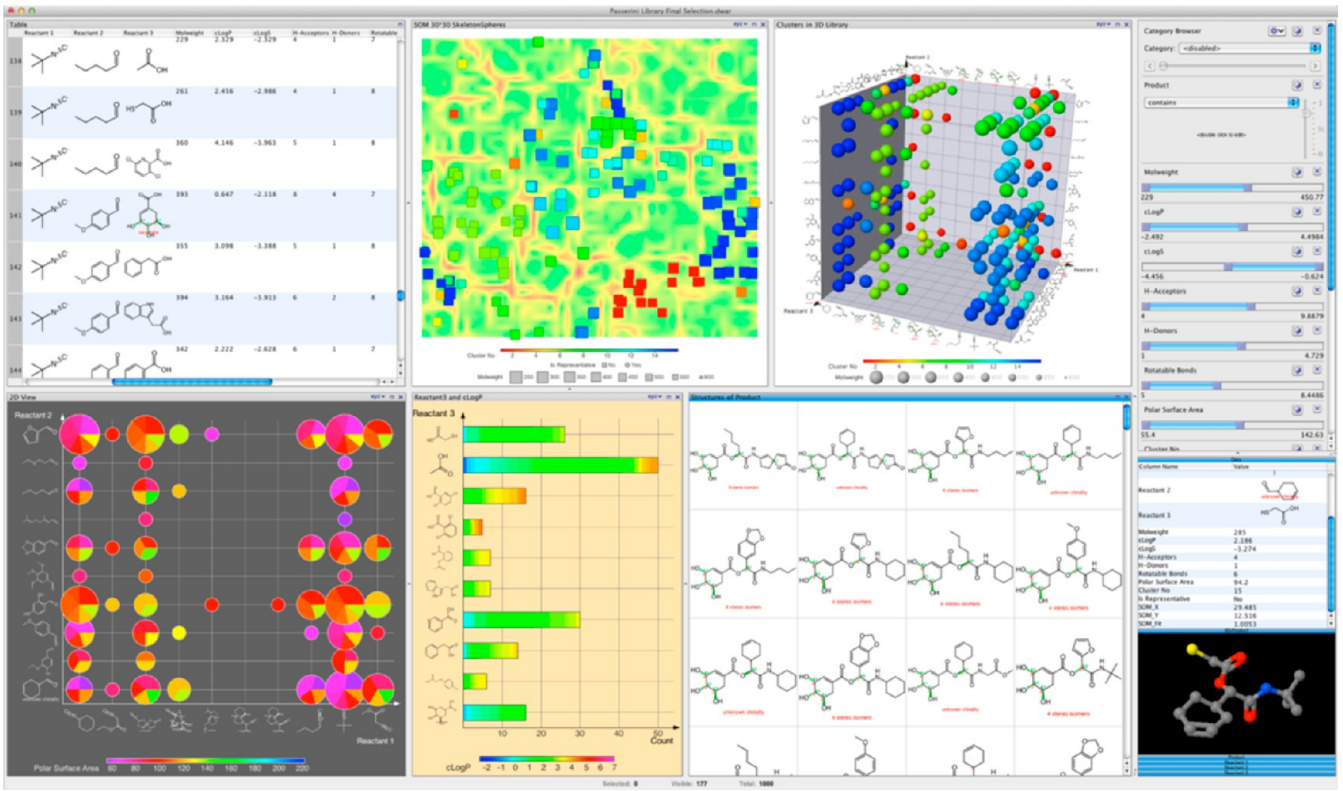


Figure 7. Combinatorial library analysis. From a generic three-component reaction and provided reactant structures, DataWarrior calculated the product structures (bottom half-right) and their physicochemical properties. Products were clustered and arranged on a 2-dimensional self-organized map (top center). Filters were adjusted (top right) and views created to show various aspects of the selected library subset.

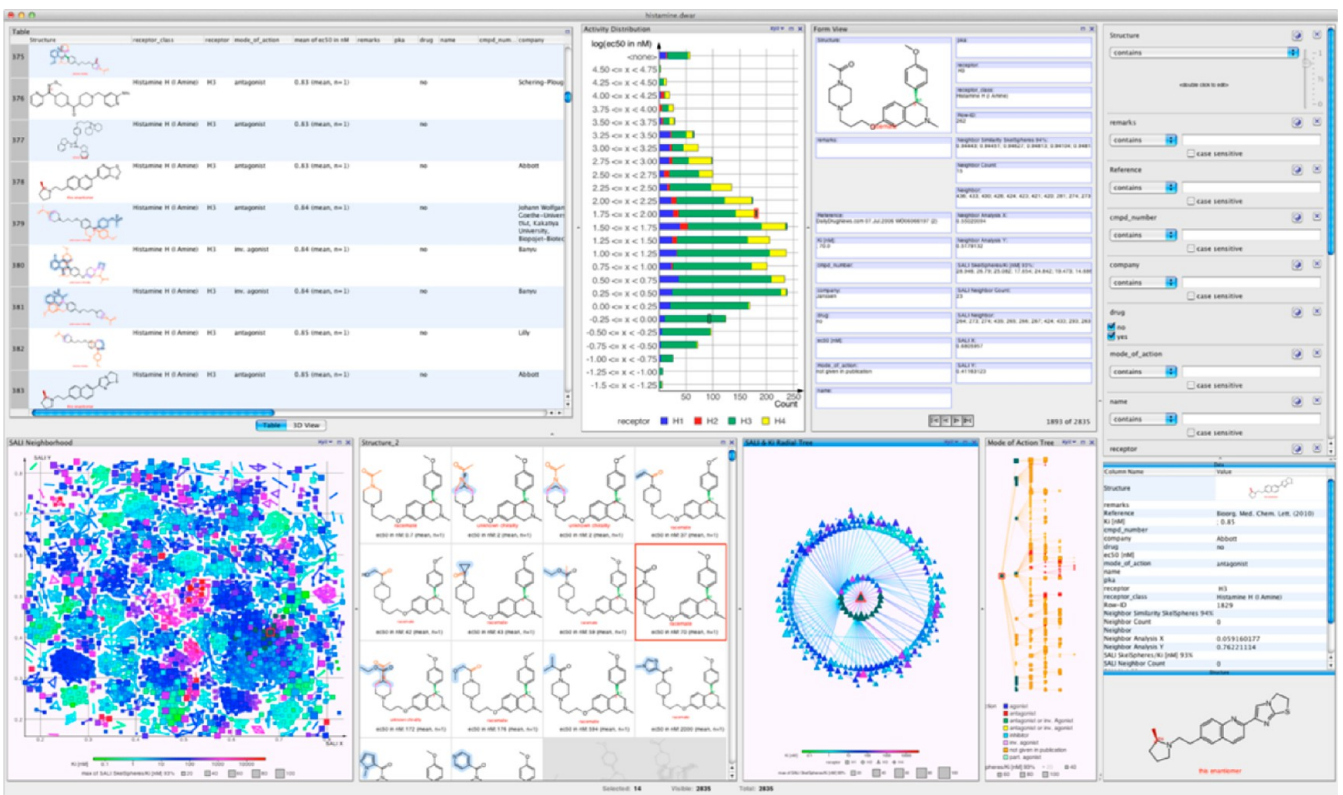


Figure 8. Activity cliff analysis of histamine receptor antagonists. Compounds are arranged in 2D by their similarity relationships. Large markers indicate activity cliffs (bottom left). Two vicinity trees with a chosen root compound show different aspects of the root's neighborhood (bottom half-right). Structural differences from the root to neighbor molecules are detected by a maximum common subgraph analysis and highlighted (bottom center).

DOI: 10.1021/ci500588j
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

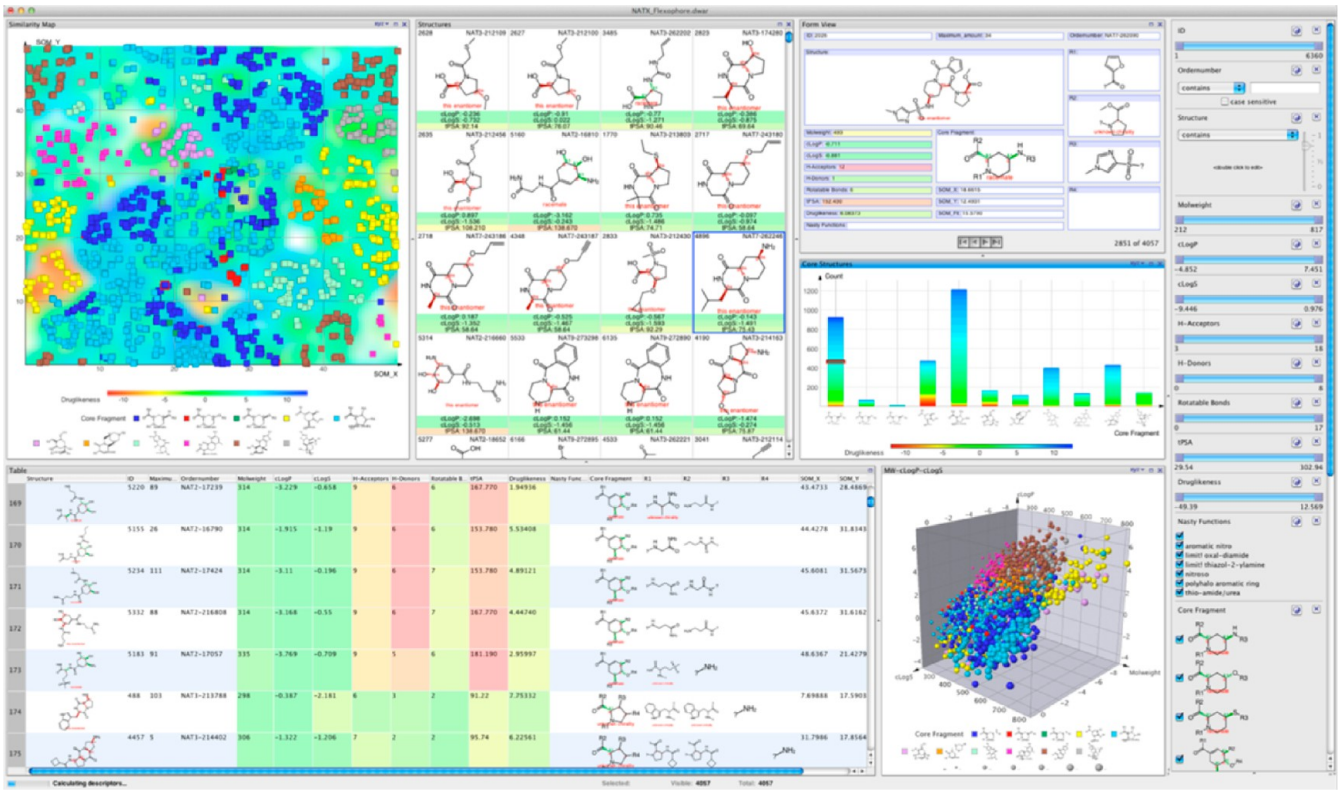


Figure 9. Scaffold analysis of a commercial natural product like library. A semiautomatic scaffold analysis perceived core fragments and substituents (bottom center). Compound properties were predicted and table cell background colors set to represent these property values applying a user-defined color scheme. Graphical views show scaffold groups on a SOM (top left) and in physicochemical property space (bottom right).

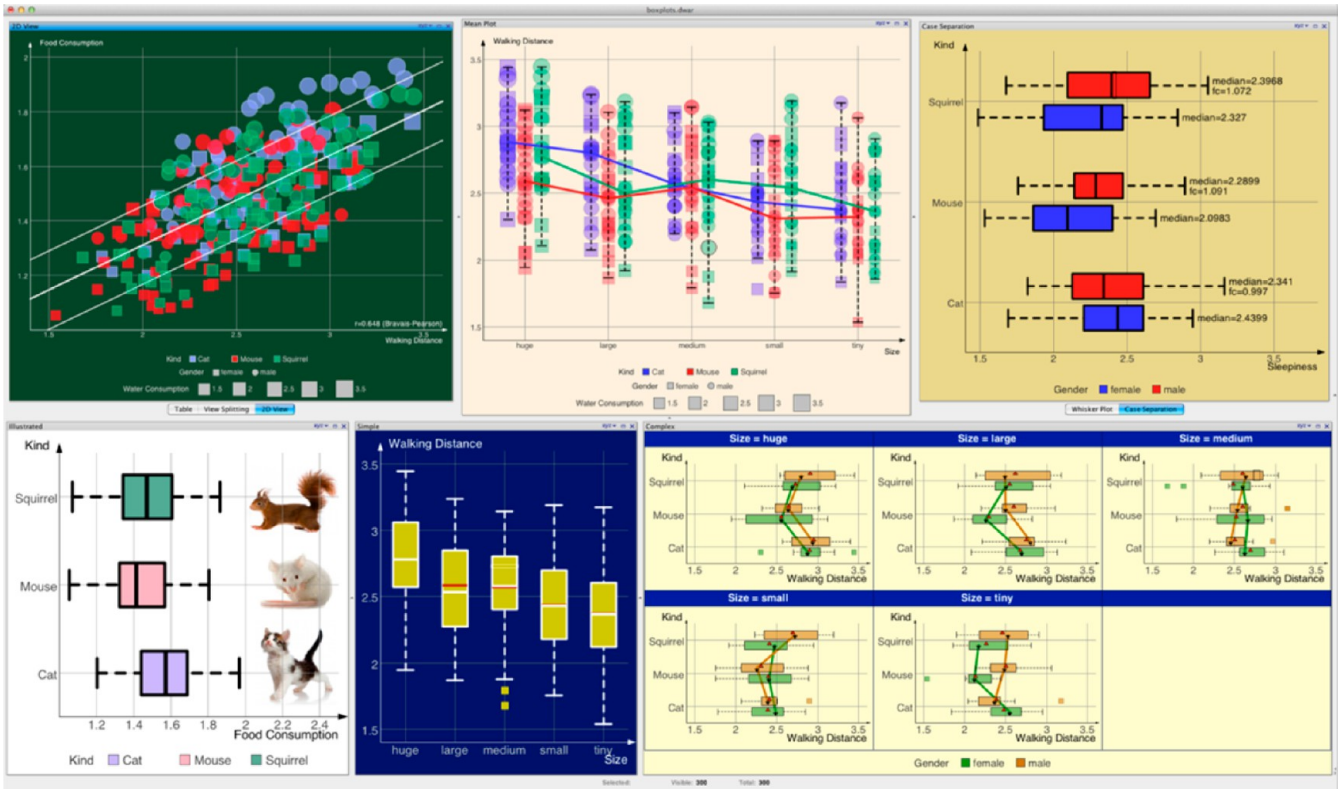


Figure 10. Correlation plot, box plots, whisker plot, and statistical parameters.

DataWarrior's Activity Cliff Analysis uses a unique 2D-scaling algorithm to topologically arrange all compounds on a 2D-view such that the most similar counterparts of any compound end up as close neighbors in the view. The algorithm is explained in

DOI: 10.1021/ci500588j
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

detail and compared to other methods in the Algorithms section. From any pair of structurally similar compounds, DataWarrior calculates the so-called structure activity landscape index[25] (SALI), which is the difference in pK_i or pIC_{50} divided by the dissimilarity. Compound activities and SALI values are represented in the view by marker color and size, respectively.

During the Activity Cliff Analysis DataWarrior locates for every compound its most similar siblings as the compound's direct neighbors. A hierarchical tree view is used to visualize a compound's neighborhood in multiple levels with the first level showing the direct neighbors, the second level showing neighbors of the direct neighbors, and so on. Such vicinity trees can be shown with radial topology or as horizontal or vertical tree. If the user clicks on a compound in any view, the vicinity tree is automatically recreated taking the new compound as a root. After an Activity Cliff Analysis, DataWarrior creates a structure view as a third view, which is configured to show selected compounds on top and to highlight structural differences between any selected compound and the root compound (Figure 8).

Scaffold Analysis. Data sets with chemical structures are often based on a limited number of reccurring scaffolds, which are decorated by changing substituents. To correlate substitution patterns with measured compound activities, two table analysis modules, one automatic and one semiautomatic, determine available scaffolds of a data set, their substitution patterns and their substituents. New columns are created that contain the scaffolds and substituents as chemical entities. These can be used in views, for filtering, or any kind of processing. Figure 9 shows a DataWarrior window with various customized views after performing a scaffold analysis of a commercially available library with allegedly natural-product-like compounds. All of the library's compounds were derived from one of 11 distinct scaffolds. In this semiautomatic analysis these 11 scaffolds were manually drawn in a structure editor. Then DataWarrior determined the substitution positions and created new columns for the scaffold and substituents.

Biological Result Analysis. In later stages of drug discovery projects, the focus shifts from many compounds with few data to few compounds with much data. Measured data of one study is often split into groups by different treatments, doses, or other parameters. Statistical parameters like r- and p-values are used to judge result significance. DataWarrior's box and whisker plots are optimized for this kind of data (Figure 10).

■ ALGORITHMS

Structure And Stereo Representation. The DataWarrior file format embeds chemical structures as compact, canonical text strings called *ID-code*, which conceptually are related to the SEMA[26,27] format (*stereo-enhanced Morgan-algorithm*) that MDL was using in their MACCS, REACCS, and Isis databases. ID-codes differ from the SEMA format in the following aspects: For molecules with multiple stereo centers or double bonds the SEMA format encoded one stereoisomer only. ID-codes are based on the enhanced stereochemical representation concept[28] suggested by MDL in 2003. This concept allows to assign one or more stereo centers to a group with defined relative configuration. Groups either belong into the *AND* or the *OR* class. In AND groups the drawn and the inverted relative configuration are present, while an OR group contains only one of those. Stereo centers that are not assigned to a group are considered absolute. This way, epimers or more complex mixtures of stereo isomers can be exactly defined within the drawing of a chemical structure. ID-codes may also represent substructures including query features in a canonical way. Moreover, ID-codes are text strings, which allow them to be included in text files and to be stored in relational database systems. Since canonical structure representations must not contain atom coordinates, an ID-code column in a DataWarrior file is typically accompanied by a second column containing the molecule's atom coordinates, which are also encoded as compact text strings. If no atom coordinates exist, these are generated on the fly whenever a molecule needs to be displayed. Since ID-codes and atom coordinate strings are very compact, native DataWarrior files have rather low disk space footprints. While the SD-file of the ChEMBL_19 database with 1 404 752 compounds requires 3.1 Gbyte, a native DataWarrior file containing the same compounds needs 318 Mbyte, which is an average of 230 bytes per compound. This not only includes the canonical structure, atom coordinates and the ChEMBL-ID, but also the *FragFp* descriptor, which is used for fast substructure and similarity search and is described in the next section.

Molecule Similarity And Molecule Descriptors. Similarities values between molecules play an important role in DataWarrior. They are used to filter compounds, to color data, to position markers, etc. Moreover, many analysis algorithms are based on compound similarities. These include clustering, self-organizing maps, an activity cliff analysis, and more. DataWarrior supports various kinds of molecule similarities ranging from a simple chemical similarity based on substructure fragments to a biological similarity that considers 3D-geometry and binding behavior.

In DataWarrior the similarity between two molecules is not calculated from a direct comparison of two chemical structures. Molecular similarities are calculated by comparing two descriptors, which were earlier derived from the molecules' structures. Both, the nature of the descriptors and the algorithm applied to compare them determines the kind of similarity value that is generated.

DataWarrior's default descriptor *FragFp* is a substructure fragment dictionary based binary fingerprint similar to the MDL keys.[29] The dictionary's 512 substructure fragments were optimized to balance high occurrence rates with low occurrence correlation in typical sets of organic compounds. The more fragments any two molecules have in common the higher is their similarity value.

The *PathFp* descriptor encodes all distinct linear strands of up to seven atoms within a molecule. Any strand is characterized by the sequence of atomic numbers and the bond orders, considering delocalized bonds as a fourth bond order. All strands are uniquely converted into a hash number, which is then used to set a corresponding bit of the fingerprint. The PathFp descriptor is related to the folded Daylight fingerprint.[30]

The *SphereFp* descriptor uniquely encodes circular fragments around every atom of the molecule. The smallest considered fragment is the naked start atom alone. Four more fragments are created from any start atom by four times adding a shell of directly connected neighbor atoms. These fragments are canonicalized and converted into a hash number to set the corresponding fingerprint bit. Circular fingerprints are similar to Hose codes[31] and were successfully applied to the prediction of ADMET properties, NMR shifts, and pK_a values.[32]

Each of the three fingerprint descriptors needs 512 bits of space only, similarities between them are quickly calculated using a Tanimoto[33] approach and they feel natural for chemists,

DOI: 10.1021/ci500588j
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

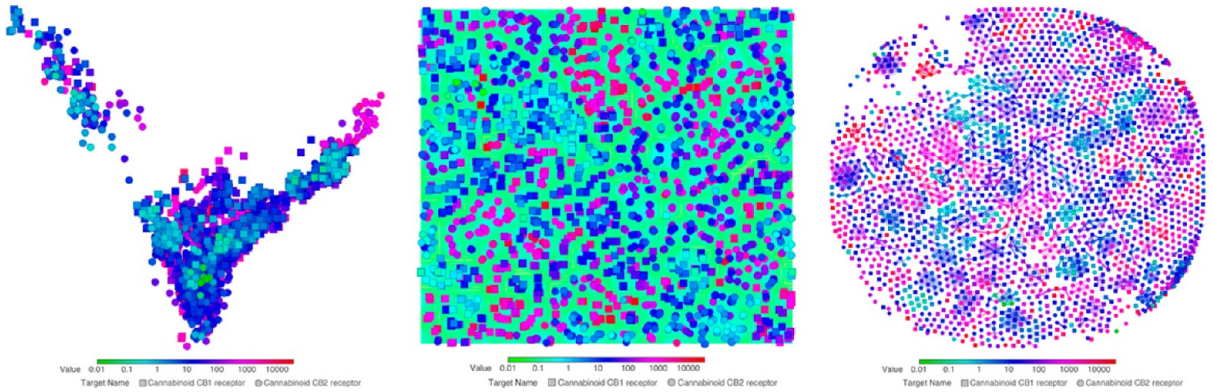


Figure 11. Chemical space visualizations of 2111 Cannabinoid receptor antagonists using three independent methods: PCA, SOM, and 2D-Rubber Band Scaling (2D-RBS). Marker colors and shapes represent activity values and receptor subtypes, respectively.

because they all encode different shades of chemical, i.e. structural similarity.

When more subtle structural changes need to perceived, the *SkeletonSpheres* descriptor should be used. It is related to the *SphereFp*, but also considers stereochemistry, counts duplicate fragments, encodes heteroatom depleted skeletons, and has twice the resolution leading to less hash collisions. On the flipside it needs more memory and similarity calculations take a little longer.

An unique *OrgFunctions* descriptor perceives molecules from a synthetic chemist's point of view, i.e. it classifies organic functional groups with a focus on steric and electronic features. 957 existing functional groups covering organic and organometallic chemistry are arranged as leaves of a binary similarity tree in a dictionary. Molecules are perceived as being similar, if they carry the same or similar functional groups in a sterically and electronically similar environment even if the synthetically inert parts of the carbon skeletons are substantially different. A detailed explanation is beyond the scope of this publication, but may be subject of a future paper.

The *Flexophore*[13] descriptor allows exploring 3D-pharmacophore similarity. It provides a very simple and yet powerful way to locate molecules with potentially similar binding behavior. Its calculation involves the creation of representative sets of conformers and may take quite some time. Different from common 3D-pharmacophore approaches, this descriptor matches entire conformer sets rather than comparing individual conformers, leading to higher predictability and taking molecular flexibility into account.

Rubber Band Scaling. One of the challenges that a medicinal chemist faces in drug discovery is to understand the nonlinear structure−property and structure−activity correlations for thousands of compounds. A frequent approach is to study long tables, which have been sorted by measured activities, by shared scaffolds or by cluster memberships. Especially, if the compounds were not synthesized as combinatorial libraries and therefore cannot be easily structured by shared scaffolds and common substituents, the approach of studying long molecule lists with experimental results from multiple assays is a rather cumbersome one.

An alternative is a chemical space visualization, where compound representing markers are placed in 2-dimensional space such that topological neighborhood corresponds with compound similarity. If in addition marker colors or sizes represent measured properties and if the software allows to interactively explore the structure and property neighborhood of individual compounds, then such a summary view may serve as an entry point to efficiently explore the entire data set. Typical methods to translate descriptor similarity into 2-dimensional space are the principal component analysis, self-organizing maps,[34] and 2-dimensional scaling.[35] DataWarrior has implementations for all three methods (Figure 11). Since DataWarrior's force field based 2-dimensional scaling approach is new and is suited for large compounds data sets, this algorithm is described in detail and compared to the PCA and SOM approach. The data set used for the comparison consisted of all selective cannabinoid receptor antagonists (CB1 and CB2) from the ChEMBL 15 database.[36]

While the PCA translates dissimilarity linearly into distance, it typically neglects more than 90% of the information on the descriptors, because the first two principal components used for the positioning often represent less than 5% of the information contained in the original multidimensional descriptor vectors. Nevertheless, usually the PCA does a surprisingly decent job of locating similar compounds closely together. Disadvantages of the PCA are that the available space is used very inefficiently and that it requires binary or numerical descriptor vectors. More complex descriptors cannot be used.

Different to the PCA, which misses most of the descriptor dimensions, the SOM neglects most of the intercompound similarity relations, because it focuses on similar compound pairs only. This is, however, not a serious disadvantage, because the goal of a chemical space visualization is that similar compounds end up as topographical neighbors and that close neighbors are always similar compounds. The SOM achieves this goal by training a 2-dimensional grid of reference vectors in such a way, that the similarity of adjacent reference vectors is maximized and that for every training compound at least one similar reference vector exists on the grid. Finally, every compound is assigned to its most similar reference vector and placed at the respective coordinates. To achieve sub grid cell resolution DataWarrior then fine-tunes compound coordinates within the cell depending on the compound's similarities to the four reference vectors of the adjacent grid cells.

The 2D-RBS focuses on similar neighborhood relations only, because in a first step it compiles all pairs of compound neighbors, whose similarity is above a certain threshold. The term *neighbor* refers to compound similarity, not to distance. The threshold is automatically determined such that on average every compound has about six neighbors. Then this list of pairs serves as the basis for the positioning algorithm. The algorithm works as follows: Initially all compounds are randomly located on a

DOI: 10.1021/ci500588j
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

quadratic space from with x and y ranging from −1.0 to 1.0. A desired final minimum distance between any two compounds is calculated as d_f = 2.0/sqrt(compound count). A compound specific neighbor factor is calculated as n_i = 4/max(4, neighbor count). Then 20 000 optimization cycles are performed for which a progress coefficient p is defined that linearly increases from 0.0 in the first cycle to 1.0 in the last cycle.

Within every cycle these values are calculated: a cycle specific collision distance $d_c = p(2 - p)d_f$, an attraction factor $a = 0.8(1 - p)\max(0.5, 1 - p)$, a repulsion factor $r = \min(0.5, 1 - p)$. For every compound i the following calculations are done: for every neighbor j the distance $d_j = \mathrm{sqrt}((x_j - x_i)^2 + (y_j - y_i)^2)$ and a distance correction $c_{f,j} = d_j - d_f$, for every colliding compound k that is closer than d_c the distance $d_k = \mathrm{sqrt}((x_k - x_i)^2 + (y_k - y_i)^2)$, and a distance correction $c_{c,k} = d_c - d_k$. At the end of each cycle all compounds' coordinates are adjusted by adding the sum of all correction contributions: $x_{new} = x + \mathrm{sumOverNeighbors}(an_i x_{f,j}) - \mathrm{sumOverColliders}(rx_{c,k})$ and $y_{new} = y + \mathrm{sumOverNeighbors}(an_i y_{f,j}) - \mathrm{sumOverColliders}(ry_{c,k})$ with $x_{f,j}$, $y_{f,j}$, $x_{c,k}$, and $y_{c,k}$ are the x and y components of $c_{f,j}$ and $c_{c,k}$.

After the completion of all coordinate minimization cycles the similarity neighborhood is well translated into topological neighborhood but the compound density increases from the edge to the center of the space. This effect increases with the number of compounds used and is caused by the fact that the center of multiple randomly located compounds is more likely to be close to the center than the location of one compound. Therefore, an individual compound, when being dragged toward the mean location of its neighbors, is often dragged toward the center. To neutralize this effect of central crowding the following simple correction is applied: All compounds are sorted by increasing distance to the center. Then, while keeping their direction from the center, their distance is updated as d_i = sqrt(i/compound count).

The three sections of the algorithm scale with about $O(n^2)$, $O(n)$, and $O(n)$ for finding the neighbors, positioning the molecules, and correcting the effect of central crowding, respectively.

2D-RBS Discussion. If a compound set is sufficiently diverse, then visualizing its chemical space in two dimensions inevitably goes along with the loss of most of the similarity information. Translating hundreds of descriptor dimensions into 2-dimensional space must either neglect most of the descriptor dimensionality or the majority of the similarity relationships between any two compounds. Luckily, for the purpose of chemical space visualization, it is not essential that low similarity relationships proportionally correlate with distance. Important, however, is that very similar compounds are located closely together. One might also wish for closely located compounds being always similar. This, though, contradicts with the criterion to efficiently use the available space. An implicit effect of efficiently using the space is that clusters of similar compounds get close to other clusters or singletons of dissimilar compounds. In DataWarrior SOMs and 2D-RBSs mitigate this unwanted effect by adding additional information to the drawing. SOMs are drawn with a background color landscape that visualizes the similarities between adjacent reference vectors. Yellow or even red ridges indicate an abrupt change of the similarity space, while adjacent locations in the same green valley contain rather similar compounds (see, e.g., SOM in Figure 7). An 2D-RBS not only aims for using the space efficiently, it distributes compounds evenly by keeping a minimum distance to adjacent compounds. While this creates a clean view where every compound can be individually recognized, compound similarities cannot be deduced from marker adjacency anymore. To make up for this, DataWarrior draws connections lines between any two similar compounds with increasing transparency reflecting decreasing similarity. This way clusters of similar compounds are easily recognized despite the equidistant compound arrangement (see e.g. Figure 8).

In order to investigate more closely, how similarity is translated into distance by the three methods, the *SkeletonSpheres* similarity and the topological distance in the view were calculated for every compound-compound combination within the data set of 2111 cannabinoid antagonists. The distribution of similarity values perceived by three different descriptors is shown in Figure 12. According to the *SkeletonSpheres* descriptor 17 777 out of

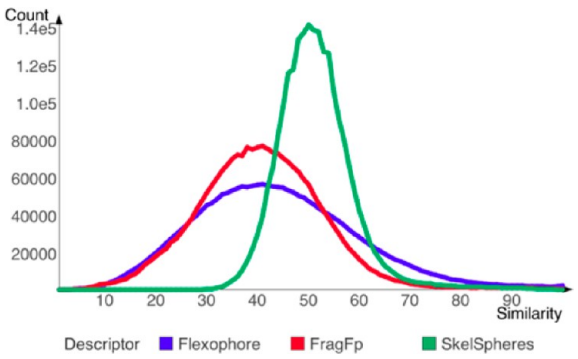


Figure 12. Distribution of pairwise similarities in 2111 cannabinoid receptor antagonists for three descriptors.

2 227 105 compound pairs have similarity values above or 80%, which about is the threshold above which a chemist considers two compounds being chemically similar.

Each of the 2 227 105 compound pairs were assigned to one of 100 similarity bins running from 0 to 100% in 1% steps. According to their topological distance they were also assigned to one of 100 distance bins. A 2-dimensional frequency plot with color encoded count values is shown in Figure 13 for PCA, SOM, and 2D-RBS on this order. We notice that any of the methods succeeds in translating similarity values above 90% into distances below 10%. As similarity values get lower toward 70% the SOM and especially the PCA manage to keep distances inversely related to similarities despite a growing dispersion. At a similarity of 50%, however, distances are broadly scattered about a mean value.

Figure 14 depicts the more relevant range above 80% similarity. All methods share the wanted effect that the maximum of the distance curve move toward higher value with decreasing similarity. But the characteristics of the three methods become evident also. In the linear PCA we have a smooth distance shift accompanies by a broadening scattering. For the SOM similarities above 94% the distance maximum is close to zero, which means that affected compounds are located at the same grid location and associated with the same reference vector. Similarities between 80 and 93% have a maximum at around 5 distance units, which is about the distance of two adjacent positions. For the 2D-RBS we have distance maxima at 2, 4, and 6 distance units, which reflect 1, 2, and 3 times the collision distance.

One big advantage of the 2D-RBS method is it applicability to nonvector descriptors. Figure 15 shows the 2D-RBS algorithm applied to 6485 serotonin receptor antagonists from the

DOI: 10.1021/ci500588j
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

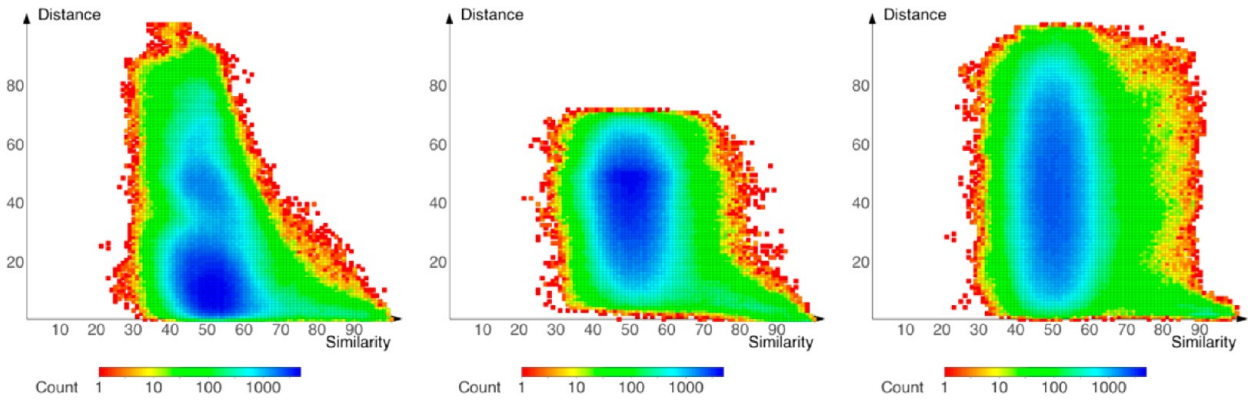


Figure 13. Two-dimensional binning of all compound pairs into similarity-distance bins for the three methods, PCA, SOM, and 2D-RBS (data set: 2111 cannabinoid receptor antagonists).

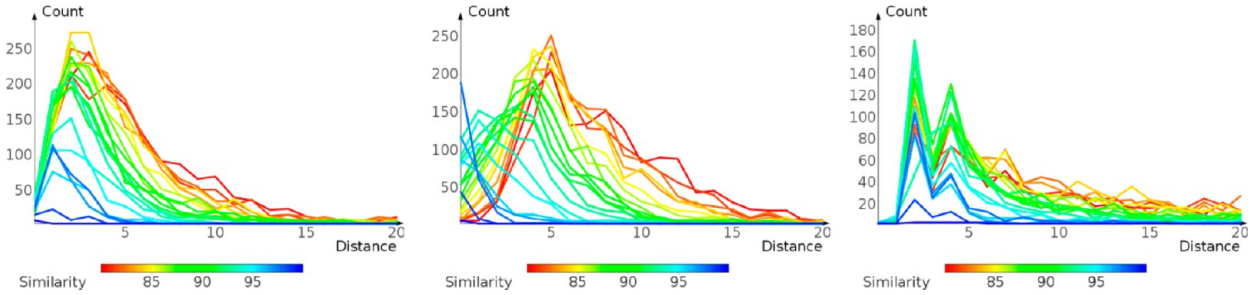


Figure 14. Translation of high similarity into distance for the three methods PCA, SOM, and 2D-RBS (data set: 2111 cannabinoid receptor antagonists).

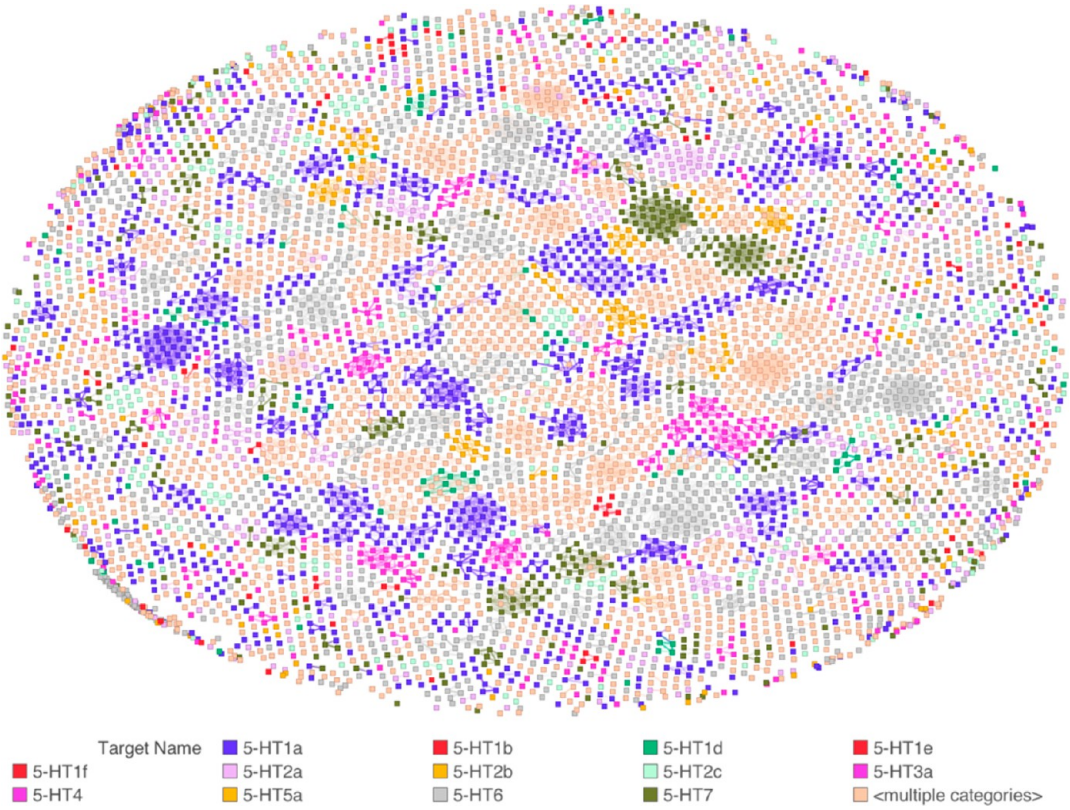


Figure 15. Flexophore space of 6485 serotonin antagonists colored by receptor subtype.

ChEMBL database using our Flexophore descriptor. We notice that the combination of the rubber band scaling with the pharmacophore similarity measure well separated compounds being active on different receptor subtypes.

DOI: 10.1021/ci500588j
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 1. Execution Times Needed for Some DataWarrior Tasks

task description	time	comment
open chembl_19.sdf file[37] with 1 404 752 compounds reading and converting all compounds into the canonical DataWarrior representation	265 s	single core
calculating the *FragFp* fingerpring for these 1 404 752 compounds in the background	395 s	
save/open native DataWarrior file with 1 404 752 compounds	17 s/11 s	single core
select 100 000 most diverse compounds from 1 404 752 compounds using *FragFp* descriptor	310 s	single core
calculate cLogP, cLogS, molecular weight, tPSA, H-donor, and H-acceptor counts for 1 404 752 compounds	48 s	
substructure search on 1 404 752 compounds using filter: m-pyridyl-C (21 475 matches)/MeO-benzyl-C (36 174 matches)	4 s/2 s	
similarity search on 1 404 752 compounds using using the default *FragFp* descriptor	<1 s	
2D-RBS scaling from 100 000 (diverse subset from above), *FragFp* descriptor similarity	2880 s	
2D-RBS scaling from 6075 Serotonine antagonists from the ChEMBL 17 database, *SkeletonSpheres* descriptor similarity	173 s	
2D-RBS scaling from 2000 Serotonine antagonists (diverse subset from above), *SkeletonSpheres* descriptor similarity	29 s	
SOM calculation from 6075 Serotonine antagonists from the ChEMBL 17 database, 100 × 100 nodes, *SkeletonSpheres* descriptor similarity	1070 s	
SOM calculation from 2000 Serotonine antagonists (diverse subset from above), 50 × 50 nodes, *SkeletonSpheres* descriptor similarity	56 s	

PEFORMANCE

Algorithms working with chemical structures are often computationally demanding. Therefore, most of DataWarrior's algorithms run on all available cores in parallel. Exceptions are reading or writing files and creating or updating views, which for technical reasons run on a single core only. Table 1 gives the completion times for various representative tasks performed by DataWarrior on a desktop computer equipped with an Intel i7-3770 quad-core CPU running Ubuntu 14.04 LTS.

When DataWarrior opens a file then the entire file content is read into memory. This design was chosen to minimize user interface delays when filter settings are changed and the focus is shifted to a different subset within a large data file. Therefore, the maximum file size that DataWarrior can open depends on the amount of memory the DataWarrior process is allowed to allocate. DataWarrior uses the Java virtual machine option −Xmx to define the maximum allocatable memory. If this parameter is defined higher than the physical free memory, then Java may refuse launching DataWarrior. Therefore, the default DataWarrior −Xmx settings are moderately set to 1.2, 3.6, 2, and 4 GB for the 32-bit Windows, 64-bit Windows, Linux, and Mac OSX versions, respectively. The memory usage per structure varies with molecule sizes, the number of additional columns, and the number of open views. However, as a rule of thumb 1 GB of memory is usually sufficient for about one million compound structures including a few columns of data and some open views.

SUMMARY

We have shown that DataWarrior is a universal data analysis and visualization program, whose embedded cheminformatics algorithms make it a versatile tool to explore large data sets of chemical structures with alphanumerical properties. DataWarrior uses both published and new cheminformatics algorithms to support synthetic and medicinal chemists in their daily work. These include combinatorial library enumeration, the prediction of molecular properties, and various methods to visualize chemical space and activity cliffs with the intent to support chemists taking smarter decisions about structural changes toward better property profiles.

A new 2-dimensional scaling method "Rubber Band Scaling" was introduced and compared to the principal component analysis and the self-organizing map for the purpose of visualizing the chemical space of medium sized compound collections. One advantage of the method is that all molecule markers are evenly distributed, and none of the molecules are hidden behind others. Hence, this method is well suited when all molecules need to be shown, e.g. for visualizing activity space or activity cliffs. Different to PCAs and SOMs, this method can be used with complex, nonvector descriptors like the *Flexophore*, which allows to visualize potential binding behavior space of chemical libraries. The *Rubber Band Scaling* algorithm is part of the DataWarrior application.

DataWarrior was released in 2014 as a free tool that can be downloaded for Linux, Macintosh or Windows from http://www.openmolecules.org/datawarrior.html.

Associated Content. The DataWarrior program for Linux, Macintosh, and Windows including full documentation, example files, and the complete Java source code under the GNU Public License are available free of charge via the Internet at http://www.openmolecules.org.

AUTHOR INFORMATION

Corresponding Author

*Phone: +41 61 565 65 23. Fax: +41 61 565 65 00. E-mail: thomas.sander@actelion.com.

Author Contributions

J.F. developed the forcefield and 3D-molecule viewer. M.v.K. developed the Flexophore descriptor. C.R. developed the PathFp fingerprint, clipboard support, and the executable wrapper for the MS-Windows operating system. T.S. developed most cheminformatics algorithms and the DataWarrior application and wrote the manuscript.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank our drug discovery management team Beat Steiner, Thomas Weller, and Martine Clozel and our CEO Jean-Paul Clozel for their continued trust and support. We also thank Isabelle Giraud for successfully educating and tutoring our user community and for delivering many useful enhancement requests. We thank all the developers who contributed to the open source libraries that DataWarrior makes use of.

ABBREVIATIONS

PCA, principal component analysis; SOM, self-organizing map; 2D-RBS, 2-dimensional rubber band scaling

REFERENCES

(1) *The Forth Paradigm: Data-Intensive Scientific Discovery*; Hey, T., Tansley, S., Tolle, K., Eds.; Microsoft Research, 2009; ISBN-13 9780982544204.

DOI: 10.1021/ci500588j
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

(2) Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I. H. The WEKA data mining software: An update. *SIGKDD Explor. Newsl.* **2009**, *11*, 10−18.
(3) *Scientific Visualization: Overviews, Methodologies, and Techniques*; Nielson, G. M., Hagen, H., Müller, H., Eds.; IEEE Computer Society, 1997.
(4) http://www.gapminder.org (accessed Sep 2014).
(5) Chen, W. L. Chemoinformatics: Past, present, and future. *J. Chem. Inf. Model.* **2006**, *46*, 2230−2255.
(6) Ott, M. A.; Noordik, J. H. Computer tools for reaction retrieval and synthesis planning in organic chemistry. A brief review of their history, methods, and programs. *Recl. Trav. Chim. Pays−Bas* **1992**, *111*, 239−246.
(7) Todd, M. H. Computer-aided organic synthesis. *Chem. Soc. Rev.* **2005**, *34*, 247−66.
(8) *TIBCO Spotfire*; TIBCO Software Inc.: Palo Alto, CA, USA, 1996.
(9) CambridgeSoft. http://www.cambridgesoft.com/ (accessed Aug 20, 2014).
(10) *Vortex*; Dotmatics: Bishops Stortford, Herts, UK, Nov. 2012.
(11) *ChemFinder Ultra*; CambridgeSoft: Cambridge, MA, USA, 1998.
(12) Sander, T.; Freyss, J.; Korff, M. v.; Reich, J. R.; Rufener, C. OSIRIS, an entirely in-house developed drug discovery informatics system. *J. Chem. Inf. Model.* **2009**, *49*, 232−246.
(13) Korff, M. v.; Freyss, J.; Sander, T. Flexophore, a new versatile 3D pharmacophore descriptor that considers molecular flexibility. *J. Chem. Inf. Model.* **2008**, *48*, 797−810.
(14) *pKa-Predictor Plugin*; ChemAxon Kft., Budapest, Hungary, 2013.
(15) The Apache Batik Project. http://xmlgraphics.apache.org/batik/ (accessed Sep 2014).
(16) Jep Java−Math Expression Parser. http://sourceforge.net/projects/jep/ (accessed Sep 2014).
(17) Jmol: an open-source Java viewer for chemical structures in 3D. http://jmol.sourceforge.net/ (accessed Sep 2014).
(18) Lowe, D. M.; Corbett, P. T.; Murray-Rust, P.; Glen, R. C. Chemical name to structure: OPSIN, an Open Source solution. *J. Chem. Inf. Model.* **2011**, *51*, 739−753.
(19) Ekit, a free open source Java HTML editor applet and application. http://www.hexidec.com/ekit.php (accessed Sep 2014).
(20) Sunflow: An open source rendering system for photo-realistic image synthesis. http://sunflow.sourceforge.net/ (accessed Sep 2014).
(21) DataWarrior User Manual, openmolecules.org. http://www.openmolecules.org/help/basics.html (accessed Nov 2014).
(22) Weininger, D.; SMILES, A. Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.
(23) The World Factbook. Central Intelligence Agency: USA. https://www.cia.gov/library/publications/download/download-2007/index.html (accessed Nov 2014).
(24) von Korff, M.; Sander, T. Toxicity-indicating structural patterns. *J. Chem. Inf. Model.* **2006**, *46*, 536−44.
(25) Guha, R.; van Drie, J. H. Structure−Activity Landscape Index: Identifying and quantifying activity cliffs. *J. Chem. Inf. Model.* **2008**, *48*, 646−658.
(26) Morgan, H. L. The Generation of a unique machine description for chemical structures-a technique developed at hemical Abstracts Service. *J. Chem. Doc.* **1965**, *5*, 107−113.
(27) Wipke, W. T.; Dyott, T. M. Stereochemically unique naming algorithm. *J. Am. Chem. Soc.* **1974**, *96*, 4834−4842.
(28) Accelrys' Enhanced Stereochemical Representation. Accelrys, Inc. http://accelrys.com/products/pdf/enhanced-stereochemical-representation.pdf (accessed Sep 2014).
(29) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Model.* **2002**, *42*, 1273−1280.
(30) James, C. A.; Weininger, D., *Daylight Theory Manual 4.34*; Daylight CIS, Inc., March 1994; Version 1, Chapter 5: Fingerprints−Screening and Similarity, pp 32−39.
(31) Bremser, W. Hose—a novel substructure code. *Anal. Chim. Acta* **1978**, *103*, 355−365.
(32) Glem, R. C.; Bender, A.; Arnby, C. H.; Carlsson, L.; Boyer, S.; Smith, J. Circular fingerprints: Flexible molecular descriptors with applications from physical chemistry to ADME. *IDrugs: the investigational drugs journal* **2006**, *9*, 199−204.
(33) Rogers, D. J.; Tanimoto, T. T. A Computer Program for Classifying Plants. *Science.* **1960**, *132*, 1115−1118.
(34) Kohonen, T. *Self-Organizing Maps*, 3rd ed.; Springer-Verlag: Berlin, 2001.
(35) Borg, I.; Groenen, P. *Modern Multidimensional Scaling: theory and applications*, 2nd ed.; Springer-Verlag: New York, 2005.
(36) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40*, D1100−D1107.
(37) ChEMBL 19 compound file, FTP link. ftp://ftp.ebi.ac.uk/pub/databases/chembl/ChEMBLdb/releases/chembl_19/chembl_19.sdf.gz (accessed Nov 2014).

DOI: 10.1021/ci500588j
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX